


Article

Gráfico de control T2 Hotelling para variables cualitativas

Wilson Rojas-Preciado^{1,2} , Mauricio Rojas-Campuzano³ , Purificación Galindo-Villardón² , Omar Ruiz-Barzola³ , , , ,

* Correspondence: wrojas@utmachala.edu.ec; Tel.: +593-992-83-3719

† Current address: Updated affiliation

‡ These authors contributed equally to this work.

Version November 14, 2021 submitted to Water



Simple Summary: A Simple summary goes here.

Abstract: Abstract

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Introduction

Los gráficos de control constituyen una de las herramientas más importantes para definir límites y parámetros óptimos de los procesos, así como para controlar la calidad de los productos mediante la reducción de la variabilidad. El uso de gráficos de control facilita la evaluación del comportamiento de las variables del proceso y contribuye al logro de los objetivos planificados.

La variación de los procesos se entiende como la diversidad de resultados que genera un grupo de variables de un proceso, su monitoreo es un objetivo clave del control estadístico, por lo tanto, es necesario entender los tipos y motivos de la variabilidad. Para ello es preciso registrar de manera sistemática y adecuada diferentes variables del proceso que se desea controlar, como las propiedades de los insumos, las condiciones de operación de los equipos, las competencias del personal que maneja los procesos, además de las características de los productos, la satisfacción de los usuarios, el cumplimiento de requisitos, entre otras.

El pionero del control estadístico de procesos fue Walter Shewhart. Estableció las diferencias entre la variabilidad natural o común, presente en todos los procesos, y la provocada por causas asignables o especiales, que pueden llevarlos a un estado de fuera de control. Señaló que un proceso está en control estadístico cuando trabaja sólo con causas comunes de variación. Propuso los primeros gráficos de control para variables de tipo continuo y para variables de atributos [1].

El control estadístico de procesos mediante gráficos de control permitió a las organizaciones monitorear el comportamiento de una variable a la vez, no obstante, las organizaciones requirieron, con el pasar del tiempo, el análisis de varias características de calidad de forma simultánea, abriendo la puerta al control estadístico de procesos desde una perspectiva multivariante [2]. Para facilitar el control de calidad de procesos es común el uso de gráficas de control que recolectan abundante información en diversas variables de forma simultánea, su análisis permite caracterizar los diferentes tipos de variables que afectan la calidad y explican su comportamiento a lo largo del tiempo [3].

Hay una variedad de gráficos de control de procesos desde la perspectiva multivariante, entre los clásicos están el Gráfico T^2 de Hotelling [4], el Multivariate Exponentially Weighted Moving –

MEWMA [5], el Multivariate Cumulative Sum Control Chart – MCUSUM [6]. Con el transcurso del tiempo se hicieron diversos aportes para mejorar el rendimiento de estos gráficos, entre los más destacados están Gráfico de control T^2 con tamaños de muestra adaptables [7], Gráfico de control T^2 con intervalos de muestreo variables [8], Gráfico de control T^2 con líneas de advertencia dobles [9], Gráfico de control robusto [10], Gráficos de control basados en modelos de minería de datos para procesos multivariantes y autocorrelacionados [11], Gráficos de control de calidad multivariantes con dimensión variable [12], Gráfico de control para el coeficiente de variación multivariante [13].

Además de estos gráficos de control para entornos paramétricos, se desarrollaron otros para datos numéricos y cualitativos en entornos multivariantes no paramétricos, entre ellos el Gráfico de control multivariante basado en la distancia de Gower para una combinación de variables continuas y cualitativas [14], Gráfico de control multivariante basado en la combinación de PCA para características de calidad de atributos y variables [15], Gráfico de control multivariante no paramétrico basado en la ponderación de novedad sensible a la densidad para procesos no normales [16], Gráfico de control de deméritos con clustering difuso de c-medias [17], Gráfico de control basado en ACP que utiliza máquinas de vectores de soporte para distribuciones no normales multivariadas [18], Gráfico CUSUM no paramétrico para monitorear procesos multivariados correlacionados en serie [19], Gráfico de control multivariante basado en Kernel PCA para monitorear características de calidad de atributos y variables mixtas [20], Gráfico T^2 basado en la combinación de PCA para datos continuos y cualitativos con detección de datos atípicos [21].

Como se puede observar, la literatura científica es abundante en lo referente a gráficos de control en entornos multivariantes paramétricos y no paramétricos para datos numéricos y, en los últimos años, para datos mixtos (numéricos y cualitativos), sin embargo, no se puede decir lo mismo de las publicaciones sobre gráficos de control multivariantes para datos cualitativos.

En el estudio de los procesos que se desarrollan en el entorno social-educativo y que explican el comportamiento de variables como el rendimiento académico, tasas de graduación o deserción, producción científica, porcentajes de matrícula de nuevo ingreso, entre otros, se maneja con mucha frecuencia variables cualitativas. No es que estén ausentes los datos cuantitativos, sino que, en las bases de datos que se utilizan para estos análisis, abundan las variables cualitativas nominales y ordinales sobre las de tipo numérico, algunos ejemplos de datos de los estudiantes son: sexo, lugar de procedencia, autodenominación étnica, grado académico de los padres, tipo de institución educativa de procedencia (fiscal, particular, municipal); ejemplos de datos de las instituciones son: tipo de sostenimiento económico, jornada, modalidad, campo de estudio, niveles (tecnológico, grado y postgrado), tipo de infraestructura; ejemplos asociados a datos de los profesores son: titularidad, dedicación, grado académico, grado en el escalafón, discapacidad, entre otros.

López [22] señala que al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión tratan de eliminarla combinando muchas variables observadas para quedarse con pocas variables ficticias que, aunque no observadas, sean combinación de las reales y sintetizan la mayor parte de la información contenida en sus datos. En este caso se deberá tener en cuenta el tipo de variables que maneja. Si son variables cuantitativas las técnicas que le permiten este tratamiento pueden ser el Análisis de componentes principales [23,24], el Análisis factorial [25–27], mientras que, si se trata de variables cualitativas, es recomendable la aplicación de un Análisis de correspondencias múltiple, Análisis de homogeneidad o un Análisis de Escalamiento multidimensional.

1.1. Análisis de Correspondencias

El tratamiento multivariante de variables cualitativas requiere un proceso metodológico distinto, uno de los más representativos es el Análisis de Correspondencias [28]. Según [22], este análisis implica estudios de similaridad o disimilaridad entre categorías, se debe cuantificar la diferencia o distancia entre ellas sumando las diferencias cuadráticas relativas entre las frecuencias de las distribuciones

de las variables analizadas, lo que conduce al concepto de la χ^2 . Así, el análisis de correspondencias puede considerarse como un análisis de componentes principales aplicado a variables cualitativas que, al no poder utilizar correlaciones, se basa en la distancia no euclídea de la χ^2 . En el enfoque francés del Análisis de Correspondencias, que se caracteriza por el énfasis en la geometría, el análisis de una tabla cruzada se llama análisis de correspondencia (CA) y el análisis de una colección de matrices indicadoras, se denomina análisis de correspondencia múltiple (MCA) [29]. En contextos anglosajones, el MCA es conocido como Análisis de Homogeneidad o Escalamiento Dual, especialmente en psicometría.

1.2. Análisis de Homogeneidad

El Análisis de Homogeneidad, Homogeneous Alternating Least Squares (HOMALS), es un modelo de la familia de modelos matemáticos del Escalamiento óptimo del sistema Gifi [30], el cual comprende una serie de técnicas exploratorias de análisis multivariado no lineal. Igual que el MCA, HOMALS se considera una forma de Análisis de Componentes Principales para datos cualitativos. El Análisis de Homogeneidad representa los objetos analizados mediante puntos en el modelo espacial, sus características más relevantes se presentan en las relaciones geométricas entre los puntos, para ello, es necesario la cuantificación de datos cualitativos [31]. El uso de variables cualitativas no es particularmente restrictivo, ya que una variable numérica continua se puede considerar como una variable cualitativa con un gran número de categorías. HOMALS se diferencia del MCA en que éste utiliza la función de Descomposición de valores propios mientras que el Análisis de Homogeneidad utiliza Mínimos Cuadrados Alternos, lo que se conoce en la literatura como la Solución de Homals [29].

1.3. Escalamiento multidimensional

Otra de las técnicas multivariantes para el tratamiento de variables cualitativas es el escalamiento multidimensional (EMD) [32,33]. Se define como una técnica que elabora una representación gráfica que permite conocer la imagen que los individuos crean de un conjunto de objetos por posicionamiento de cada uno en relación con los demás, (mapa perceptual). El EMD trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual, normalmente de dos o tres dimensiones, de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas. Las dimensiones de este espacio conceptual se pueden interpretar para favorecer la comprensión de los datos, inclusive si son valoraciones subjetivas de disimilaridad entre objetos o conceptos. Dado que el EMD permite calcular las distancias a partir de los datos multivariados, si las variables se han medido objetivamente, se lo puede utilizar como técnica de reducción de datos [22].

Las variables que utiliza esta técnica pueden ser métricas o no métricas. El escalamiento multidimensional las transforma en distancias entre los objetos en un espacio de dimensiones múltiples, de modo que objetos que aparecen situados más próximos entre sí son percibidos como más similares por los sujetos.

2. Materials and Methods

2.1. Notación

La tabla 1 contiene elementos, representación y ejemplos de la manera como se presentan los elementos algebraicos abordados en la metodología.

A lo largo del artículo se utilizarán letras para hacer referencia a parámetros necesarios, se los enuncia a continuación en la tabla 2:

Elementos	Representación	Ejemplo
Escalares	Letras en minúscula.	v, λ
Vectores	Letras en minúscula y en negrita.	\mathbf{v}, \mathbf{u}
Matrices	Letras en mayúscula y en negrita.	\mathbf{V}, \mathbf{X}
Matrices de tres vías (Cubos de datos)	Letras con doble trazo en mayúscula.	\mathbb{C}, \mathbb{X}

Table 1. Elementos algebraicos

Letra	Significado	Especificación
p	Número de dimensiones	
K	Número total de tablas (Especifica la profundidad del cubo de datos)	
k	Índice de tabla	$k=1,2,\dots,K$
T	Índice de matriz transpuesta	\mathbf{X}^T
n	Tamaño muestral de las K tablas	

Table 2. Notación

2.2. Análisis de Correspondencia Múltiple (MCA)

El análisis de correspondencias múltiples (MCA) es una generalización del análisis de correspondencias simple o binario, donde se incluyen más variables cualitativas. Se obtiene al realizar el análisis de correspondencias simple a una tabla disyuntiva completa, conocida como la tabla de Burt.

Se tiene una matriz de datos con p variables cualitativas, cada una con h categorías ($h > 1$). En el ejemplo que se desarrolla para esta investigación, se dispone de una base de datos (*Data10Contaminated*) constituida por 10 tablas, cada una tiene 10 variables y cada variable, 3 categorías (Alto, Medio y Bajo).

V_1	V_2	\dots	V_p
Alto	Medio	\dots	Medio
Medio	Bajo	\dots	Alto
\vdots	\vdots	\vdots	\vdots
Bajo	Alto	\dots	Bajo

Table 3. Matriz inicial

Esta matriz es equivalente a la matriz disyuntiva Z , que desglosa las variables en cada una de sus modalidades y se registra la ocurrencia de eventos de forma binaria.

V_1 Alto	V_1 Medio	V_1 Bajo	V_2 Alto	V_2 Medio	V_2 Bajo	\dots	V_p Alto	V_p Medio	V_p Bajo
1	0	0	1	0	0	\dots	0	1	0
0	1	0	0	1	0	\dots	1	0	0
0	0	1	0	0	1	\dots	0	0	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
1	0	0	1	0	0	\dots	1	0	0

Table 4. Matriz disyuntiva Z

La tabla de Burt viene dada por:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \quad (1)$$

La construcción de la matriz de Burt se da por la superposición de tablas. En las tablas ubicadas en la diagonal se encuentran matrices diagonales que contienen las frecuencias marginales de cada una de las variables. Fuera de la diagonal de la matriz de Burt se encuentran las tablas cruzadas por pares de variables.

Para realizar el análisis de correspondencias múltiples se parte de la matriz de Burt, obtenida con la ecuación 1. Esta matriz está formada por las frecuencias absolutas, éstas se transforman en frecuencias relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a una nueva matriz que se denominará **P**.

	V_1 Alto	V_1 Medio	V_1 Bajo	V_2 Alto	V_2 Medio	V_2 Bajo	...	V_p Alto	V_p Medio	V_p Bajo
V_1 : Alto	$b_{1,1}$	0	0	$b_{1,4}$	$b_{1,5}$	$b_{1,6}$...	$b_{1,3p-2}$	$b_{1,3p-1}$	$b_{1,3p}$
V_1 : Medio	0	$b_{2,2}$	0	$b_{2,4}$	$b_{2,5}$	$b_{2,6}$...	$b_{2,3p-2}$	$b_{2,3p-1}$	$b_{2,3p}$
V_1 : Bajo	0	0	$b_{3,3}$	$b_{3,4}$	$b_{3,5}$	$b_{3,6}$...	$b_{3,3p-2}$	$b_{3,3p-1}$	$b_{3,3p}$
V_2 : Alto	$b_{4,1}$	$b_{4,2}$	$b_{4,3}$	$b_{4,4}$	0	0	...	$b_{4,3p-2}$	$b_{4,3p-1}$	$b_{4,3p}$
V_2 : Medio	$b_{5,1}$	$b_{5,2}$	$b_{5,3}$	0	$b_{5,5}$	0	...	$b_{5,3p-2}$	$b_{5,3p-1}$	$b_{5,3p}$
V_2 : Bajo	$b_{6,1}$	$b_{6,2}$	$b_{6,3}$	0	0	$b_{6,6}$...	$b_{6,3p-2}$	$b_{6,3p-1}$	$b_{6,3p}$
...
V_p : Alto	$b_{3p-2,1}$	$b_{3p-2,2}$	$b_{3p-2,3}$	$b_{3p-2,4}$	$b_{3p-2,5}$	$b_{3p-2,6}$...	$b_{3p-2,3p-2}$	0	0
V_p : Medio	$b_{3p-1,1}$	$b_{3p-1,2}$	$b_{3p-1,3}$	$b_{3p-1,4}$	$b_{3p-1,5}$	$b_{3p-1,6}$...	0	$b_{3p-1,3p-1}$	0
V_p : Bajo	$b_{3p,1}$	$b_{3p,2}$	$b_{3p,3}$	$b_{3p,4}$	$b_{3p,5}$	$b_{3p,6}$...	0	0	$b_{3p,3p}$

Table 5. P: Tabla de contingencia de Burt en frecuencias relativas

Se obtienen las marginales de las filas (*mf*) y de las columnas (*mc*) de la matriz **P** (Tabla 5). A estos vectores se los conoce también como *Masas de fila y columna*, respectivamente.

V_1 Alto	V_1 Medio	V_1 Bajo	V_2 Alto	V_2 Medio	V_2 Bajo	...	V_p Alto	V_p Medio	V_p Bajo
$b_{\bullet,1}$	$b_{\bullet,2}$	$b_{\bullet,3}$	$b_{\bullet,4}$	$b_{\bullet,5}$	$b_{\bullet,6}$...	$b_{\bullet,3p-2}$	$b_{\bullet,3p-1}$	$b_{\bullet,3p}$

Table 6. Frecuencias marginales de las filas. (mf)

V_1 Alto	V_1 Medio	V_1 Bajo	V_2 Alto	V_2 Medio	V_2 Bajo	...	V_p Alto	V_p Medio	V_p Bajo
$b_{\bullet,1}$	$b_{\bullet,2}$	$b_{\bullet,3}$	$b_{\bullet,4}$	$b_{\bullet,5}$	$b_{\bullet,6}$...	$b_{\bullet,3p-2}$	$b_{\bullet,3p-1}$	$b_{\bullet,3p}$

Table 7. Frecuencias marginales de las columnas. (mc)

Se obtiene la matriz de residuos estandarizados **S**.

$$\mathbf{S} = \mathbf{D}_{\text{fila}}^{-\frac{1}{2}} (\mathbf{P} - \mathbf{mf} \mathbf{mc}') \mathbf{D}_{\text{columna}}^{-\frac{1}{2}} \quad (2)$$

donde:

- \mathbf{D}_{fila} es una matriz diagonal que contiene las masas de las filas.
- $\mathbf{D}_{\text{columna}}$ es una matriz diagonal que contiene las masas de las columnas

Se aplica descomposición singular (SVD) a la matriz **S** (Ecuación 2):

$$\mathbf{S} = \mathbf{UDV}' \quad (3)$$

donde:

- \mathbf{U} y \mathbf{V} son matrices ortogonales.
- \mathbf{D} es una matriz diagonal que contiene los valores singulares.

Para encontrar las coordenadas estandarizadas se aplica lo siguiente:

$$\mathbf{X} = \mathbf{D}_{\text{fila}}^{-\frac{1}{2}} \mathbf{U} \quad (4)$$

$$\mathbf{Y} = \mathbf{D}_{\text{columna}}^{-\frac{1}{2}} \mathbf{V} \quad (5)$$

Para los fines necesarios, se utilizará las coordenadas de las columnas (Tabla 8).

	Dim_1	Dim_2	\dots	Dim_{3p}
$V_1 : Alto$	$v_1 d_{1alto}$	$v_1 d_{1alto}$	\dots	$v_1 d_{palto}$
$V_1 : Medio$	$v_1 d_{1medio}$	$v_1 d_{1medio}$	\dots	$v_1 d_{pmedio}$
$V_1 : Bajo$	$v_1 d_{1bajo}$	$v_1 d_{1bajo}$	\dots	$v_1 d_{pbajo}$
\vdots	\vdots	\vdots	\ddots	\vdots
$V_p : Bajo$	$v_p d_{1bajo}$	$v_p d_{1bajo}$	\dots	$v_p d_{pbajo}$

Table 8. Coordenadas estandarizadas de las columnas.

2.3. Generalización a K tablas

Si se tienen K tablas, con la misma estructura de la tabla 3, como se visualiza en la figura 1, se aborda el enfoque del análisis factorial múltiple (MFA). Escofier and Pagès [34] indica que el MFA utiliza análisis de correspondencia múltiple cuando se trata de variables cualitativas. El procedimiento implica la realización de un MCA por cada tabla y dividirlo por su primer valor propio con la finalidad de obtener K grupos normalizados. Posteriormente se consideran todas las tablas y se realiza un MCA global.

	V_1	V_2	\dots	V_p
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				
Medio				
Medio				
V_1	V_2	\dots	V_p	
Alto				

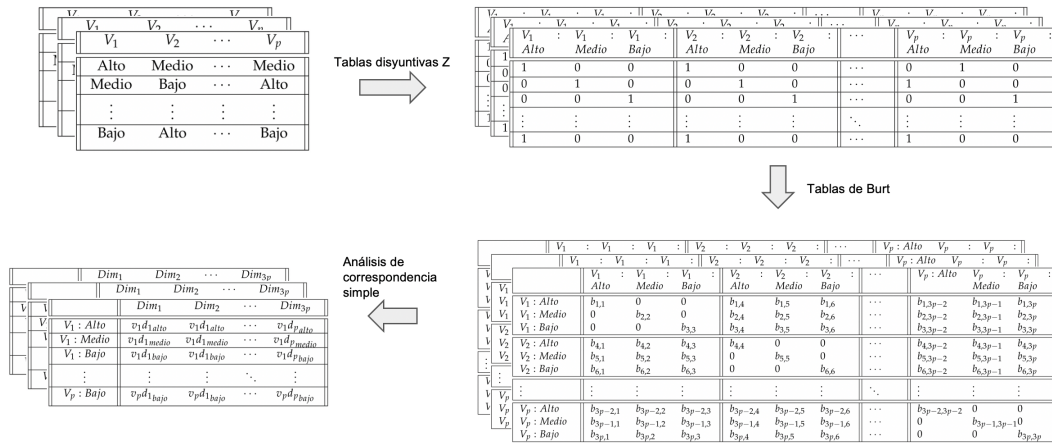


Figure 2. Procedimiento del MCA para K tablas

Sea λ_1^k el primer valor propio obtenido de la descomposición singular de la k-ésima tabla C. Se normaliza la tabla multiplicándola por $1/\lambda_1^k$. Con esto se obtiene la tabla C' , que corresponde a la tabla de coordenadas normalizadas.

Individualmente, para el caso de la matriz k, se tendría la siguiente expresión.

$$C'_k = \frac{1}{\lambda_1^k} C_k \quad (6)$$

Aglomerando las matrices normalizadas C' en una sola, se tiene la matriz C' . Esta contiene todos los elementos de las k tablas.

$$C' = [C'_1 | C'_2 | \dots | C'_k]^T \quad (7)$$

La normalización que realiza el MFA se encarga de ponderar las k tablas, con el objetivo de evitar alguna descompensación al momento de realizar el análisis conjunto de las tablas.

2.5. Gráfico de control

Para definir el gráfico de control T^2 Hotelling se deben tomar las siguientes consideraciones:

- La tabla C' (Ecuación 7) se denomina Consenso, sirve como referente para el escenario *bajo control*, y de la cual se obtiene μ_0 y S_0 .
- Cada matriz C'_k tiene el mismo número de filas (n) y columnas (p) (individuos y variables).
- El vector de medias $\bar{\mu}_k$ está atado a la tabla C'_k , es decir, el gráfico de control estará en función de las diferencias entre las matrices C'_k y la matriz consenso C' .
- Las matrices C'_k siguen una distribución normal multivariante con vector de medias μ_k y matriz de covarianzas S_k .

Con esto se obtiene el estadístico T^2 :

$$T^2 = n(\mu_k - \mu_0)' \Sigma_0^{-1} (\mu_k - \mu_0) \quad (8)$$

Se sabe que, bajo control, el T^2 se distribuye como una Chi-cuadrado con p grados de libertad χ_p^2 . En este caso se puede aplicar este principio, ya que se utiliza la matriz consenso (C'), que representa al escenario bajo control.

Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, sólo tiene límite de control superior. Este viene dado por la ecuación 9

$$UCL = \chi^2_{\alpha,p} \quad (9)$$

donde p es el número de dimensiones y α es la significancia predeterminada considerando p .

2.6. Tabla posterior

Con la finalidad de detectar las potenciales categorías responsables de que un punto en el gráfico T^2 de Hotelling para variables cualitativas se encuentre fuera de control, se propone una tabla que presenta las anomalías de cada categoría en cada variable, comparando las masas de columna de la tabla k y las masas de columna de la tabla consenso por medio de distancias χ^2 que proporcionan un valor p , aportando a la interpretación.

3. Complemento computacional

Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete reproducible en R. El paquete **T2Qv** utiliza la metodología expuesta en este artículo y la lleva a un entorno práctico, permite visualizar los resultados de forma plana o interactiva, además, presenta un panel Shiny que contiene todas las funciones individuales en un mismo espacio.

3.1. Disponibilidad

El paquete está disponible en GitHub, la descarga se la puede realizar de la siguiente forma:

```
install.packages("devtools")
devtools::install_github("JavierRojasC/T2Qv")
```

3.2. El paquete: T2Qv

```
Package: T2Qv
Type: Package
Title: Control Qualitative Variables
Version: 0.1.0
Authors@R: c(person("Wilson", "Rojas-Preciado", role = c("aut", "cre"),
  email = "wrojas@utmachala.edu.ec"),
  person("Mauricio", "Rojas-Campuzano", role = c("aut", "ctb"),
  email = "mauroja@espol.edu.ec"),
  person("Purificación", "Galindo-Villardón", role = c("aut", "ctb"),
  email = "oruiz@espol.edu.ec"),
  person("Omar", "Ruiz-Barzola", role = c("aut", "ctb"),
  email = "oruiz@espol.edu.ec"))
Maintainer: Wilson Rojas-Preciado <wrojas@utmachala.edu.ec>
Description: Covers k-table control analysis using multivariate control charts for qualitative variables using
fundamentals of multiple correspondence analysis and multiple factor analysis. The graphs can be shown in a
flat or interactive way, in the same way all the outputs can be shown in an interactive shiny panel.
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.1.1
Depends: R (>= 2.10)
Imports: shiny, shinydashboardPlus, shinydashboard, shinycssloaders,
  dplyr, ca, highcharter, stringr, tables, htmltools (>= 0.5.1.1)
Suggests: testthat (>= 3.0.0)
Config/testthat/edition: 3
Author: Wilson Rojas-Preciado [aut, cre],
  Mauricio Rojas-Campuzano [aut, ctb],
  Purificación Galindo-Villardón [aut, ctb],
  Omar Ruiz-Barzola [aut, ctb]
Built: R 4.0.2; ; 2021-10-14 23:56:56 UTC; unix
```

Figure 3. Documentación del paquete T2Qv

Las funciones que contiene el paquete y su descripción se enuncian en la tabla 9.

4. Análisis de sensibilidad

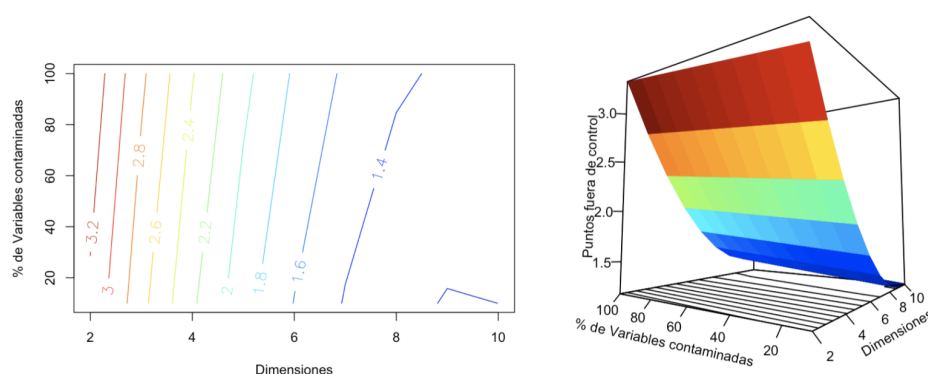
Como se ha mencionado, en el gráfico T2Qv un punto fuera de control se interpreta como una tabla (k_i) que incluye una cantidad o una proporción de variables contaminadas, de tal manera que la diferencia de los valores de masas de columna, entre de la matriz k_i y la matriz consenso, sean

Función	Descripción
T2 qualitative	Multivariate control chart T2 Hotelling applicable for qualitative variables.
MCAconsensus	Multiple correspondence analysis applied to a consensus table.
MCApoint	Multiple correspondence analysis applied to a specific table.
ChiSq variable	Contains Chi square distance between the column masses of the table specified in PointTable and the consensus table. It allows to identify which mode is responsible for the anomaly in the table in which it is located.
Full Panel	A shiny panel complete with the multivariate control chart for qualitative variables, the two MCA charts and the modality distance table. Within the dashboard, arguments such as type I error and dimensionality can be modified.

Table 9. Funciones del paquete T2Qv

significativos según el valor p obtenido de la distribución χ^2 . En estos casos, se espera que los puntos en el gráfico T2Qv generalicen el comportamiento de estas diferencias y superen el límite de control superior (UCL). La ubicación de este límite de control varía en función del número de dimensiones que se representen, así, cuando es alto se logra un desempeño óptimo, mientras que, se introduce inestabilidad y se pierde confiabilidad en los resultados al disminuir el número de dimensiones de entre las que se puede representar.

El gráfico de control propuesto es capaz de detectar un punto fuera de control, aún con un bajo número de variables contaminadas, cuando se trabaja con un alto número de dimensiones. Se recomienda $p - 1$, tal que p es el número total de dimensiones de la matriz inicial (Tabla 3). Cuando se disminuye el número de dimensiones también disminuye la altura del límite de control superior (UCL), en consecuencia, se incrementa el número de puntos fuera de control, aunque no necesariamente las variables expresen diferencias significativas en sus valores, crece la probabilidad de falsos positivos. Por consiguiente, la pregunta que surge es hasta cuántas dimensiones se puede disminuir en el análisis sin perder confiabilidad en el resultado. La importancia de esta pregunta radica en la necesidad de disponer un gráfico confiable, que identifique puntos fuera de control aún si se ha aplicado a los datos una técnica de una reducción de dimensiones, sin caer en casos de falso positivo.

**Figure 4.** Curvas de nivel y superficie de respuesta obtenidas con el gráfico T2 Hotelling para variables cualitativas.

El análisis de sensibilidad utiliza curvas de nivel y superficies de respuesta (figura 4) para representar el número de puntos fuera de control, considerando el porcentaje de variables contaminadas de la k_i tabla y el número de dimensiones representadas. Los datos de prueba utilizados en el modelo se registran en 10 tablas, cada una de ellas incluye 10 variables y cada variable tiene tres categorías: alto, medio y bajo. La tabla 10 tiene una distribución diferente de las demás, esta es la tabla contaminada. Se observa que el modelo es capaz de identificar un punto fuera de control trabajando con 9 dimensiones ($p-1$), aún con un porcentaje bajo de variables contaminadas. Cuando el número

de dimensiones disminuye a 8 y el porcentaje de variables contaminadas es cercano a 100%, detecta correctamente 1 punto fuera de control. Se observa además que cuando el número de dimensiones es menor se pierde estabilidad. En consecuencia, el análisis de sensibilidad ratifica que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones.

Appendix A

Appendix A.1

Appendix B

References

- Gutiérrez, H.; de la Vara Salazar, R. *Control estadístico de la calidad y seis sigma*; Vol. 3, McGraw Hill Education, 2013; p. 152–253.
- Ramos, M. Una alternativa a los métodos clásicos de control de procesos basada en coordenadas paralelas, métodos Biplot y Statis. PhD thesis, 2017.
- Li, J.; Tsung, F.; Zou, C. Directional control schemes for multivariate categorical processes. *Journal of Quality Technology* **2012**, *44*, 136–154.
- Hotelling, H. Multivariate quality control. Techniques of statistical analysis. McGraw-Hill, New York **1947**.
- Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving average control chart. *Technometrics* **1992**, *34*, 46–53.
- Crosier, R.B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* **1988**, *30*, 291–303.
- APARISI, F. Hotelling's T2 control chart with adaptive sample sizes. *International Journal of Production Research* **1996**, *34*, 2853–2862, [<https://doi.org/10.1080/00207549608905062>]. doi:10.1080/00207549608905062.
- Aparisi, F.; Haro, C.L. Hotelling's T2 control chart with variable sampling intervals. *International Journal of Production Research* **2001**, *39*, 3127–3140, [<https://doi.org/10.1080/00207540110054597>]. doi:10.1080/00207540110054597.
- Faraz, A.; Parsian, A. Hotelling's T2 control chart with double warning lines. *Statistical Papers* **2006**, *47*, 569–593. doi:10.1007/s00362-006-0307-x.
- Shabbak, A.; Midi, H. An improvement of the hotelling statistic in monitoring multivariate quality characteristics. *Mathematical Problems in Engineering* **2012**, *2012*.
- Kim, S.B.; Jitpitaklert, W.; Park, S.K.; Hwang, S.J. Data mining model-based control charts for multivariate and autocorrelated processes. *Expert Systems with Applications* **2012**, *39*, 2073–2081.
- Ruiz-Barzola, O. Gráficos de Control de Calidad Multivariantes con Dimension Variable. PhD thesis, Universitat Politècnica de València, 2013.
- Yeong, W.C.; Khoo, M.B.C.; Teoh, W.L.; Castagliola, P. A control chart for the multivariate coefficient of variation. *Quality and Reliability Engineering International* **2016**, *32*, 1213–1225.
- Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications* **2014**, *41*, 1701–1707. doi:10.1016/j.eswa.2013.08.068.
- Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research* **2018**, *6*, 364–384, [<https://doi.org/10.1080/21693277.2018.1517055>]. doi:10.1080/21693277.2018.1517055.
- Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
- YILMAZ, H.; Yanik, S. Design of Demerit Control Charts with Fuzzy c-Means Clustering and an Application in Textile Sector. *Textile and Apparel* **2020**, *30*, 117–125.
- Farokhnia, M.; Niaki, S.T.A. Principal component analysis-based control charts using support vector machines for multivariate non-normal distributions. *Communications in Statistics -*

- Simulation and Computation* **2020**, *49*, 1815–1838, [<https://doi.org/10.1080/03610918.2018.1506032>]. doi:10.1080/03610918.2018.1506032.
19. Xue, L.; Qiu, P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology* **2020**, pp. 1–14.
 20. Ahsan, M.; Mashuri, M.; Wibawati.; Khusna, H.; Lee, M.H. Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics. *Symmetry* **2020**, *12*. doi:10.3390/sym12111838.
 21. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Outlier detection using PCA mix based T2 control chart for continuous and categorical data. *Communications in Statistics - Simulation and Computation* **2021**, *50*, 1496–1523, [<https://doi.org/10.1080/03610918.2019.1586921>]. doi:10.1080/03610918.2019.1586921.
 22. López, C.P. *Técnicas de análisis multivariante de datos*; Pearson Educación, 2004.
 23. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series* **1901**, *2*, 417 – 441. doi:10.1080/14786440109462720.
 24. Hotelling, H. Analysis of a complex of statistical variables into principal components. **1933**. *24*, 417 – 441. doi:10.1037/h0071325.
 25. Ch, S.; others. General intelligence objectively determined and measured. *American Journal of Psychology* **1904**, *15*, 201–293.
 26. Thurstone, L.L. Multiple-factor analysis; a development and expansion of The Vectors of Mind. **1947**.
 27. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
 28. Benzecri., J. *OL'analyse des correspondances. En L'Analyse des Données: Leçons sur L'analyse Factorielle et la Reconnaissance des Formes et Travaux*; Paris - 1973, 1973.
 29. Michailidis, G.; Leeuw, J.D. The Gifi system of descriptive multivariate analysis. *Statistical Science* **1998**, pp. 307–336.
 30. Gifi, A. *Nonlinear multivariate analysis*; Vol. 14, John Wiley & Sons, 1990.
 31. López de Ipiña, F. Análisis multivariante aplicado al estudio del parentesco. Representaciones HOMALS. **2014**.
 32. Torgerson, W.S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401–419.
 33. Shepard, R.N. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* **1962**, *27*, 125–140.
 34. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis* **1994**, *18*, 121–140. doi:[https://doi.org/10.1016/0167-9473\(94\)90135-X](https://doi.org/10.1016/0167-9473(94)90135-X).